

Best reply structure and equilibrium convergence in generic games

Marco Pangallo,^{1,2,*} Torsten Heinrich,^{1,2} and J. Doyne Farmer^{1,2,3,4}

¹*Institute for New Economic Thinking at the Oxford Martin School, University of Oxford, Oxford OX2 6ED, UK*

²*Mathematical Institute, University of Oxford, Oxford OX1 3LP, UK*

³*Computer Science Department, University of Oxford, Oxford OX1 3QD, UK*

⁴*Santa Fe Institute, Santa Fe, NM 87501, US*

(Dated: April 19, 2017)

Game theory often assumes rational players that play equilibrium strategies. But when the players have to learn their strategies by playing the game repeatedly, how often do the strategies converge? We analyze generic two player games using a standard learning algorithm, and also study replicator dynamics, which is closely related. We show that the frequency with which strategies converge to a fixed point can be understood by analyzing the best reply structure of the payoff matrix. A Boolean transformation of the payoff matrix, replacing all best replies by one and all other entries by zero, provides a reasonable approximation of the asymptotic strategic dynamics. We analyze the generic structure of randomly generated payoff matrices using combinatorial methods to compute the frequency of cycles of different lengths under the microcanonical ensemble. For a game with N possible moves the frequency of cycles and non-convergence increases with N , becoming dominant when $N > 10$. This is especially the case when the interactions are competitive.

Classical game theory is the study of strategic interactions between individual players [1], but its tools have also proven useful to model evolutionary processes [2, 3], social phenomena such as the emergence of cooperation [4], language formation [5] and opinion dynamics [6]. The typical approach in game theory is to solve the resulting game by assuming that the players instantly coordinate on an equilibrium. But in a context where players are not fully rational and must learn their strategies by playing the game repeatedly, they may fail to converge. How often does this happen?

Most of the existing literature addressing this question has focused on specific games or classes of games [7–17]. More recently, work by statistical physicists has addressed a very general ensemble of games, but has not provided sufficient understanding of what determines the stability of equilibria. Letting N be the number of moves, the asymptotic behavior of these games in the limit $N \rightarrow \infty$ has been characterized using the replica trick or path integral methods, both developed to study spin-glasses [18–21]. This is a good approximation for $N > 50$, but it leaves open the question of what happens for smaller N . Our main contribution here is to introduce a new method to estimate the typical behavior for any N , that also provides insight into the underlying behavior of learning algorithms and evolutionary processes. This method depends on the best reply structure of the payoff matrix, a basic property of the game that was never systematically investigated before. Given any game, it provides an approximate estimate of the probability of non-convergence.

Assume a two player game in which player Row chooses move $i = 1, \dots, N$ with probability $x_i(t)$ and player Column chooses move $j = 1, \dots, N$ with probability $y_j(t)$. We study Experience-Weighted Attraction (EWA), a generalization of reinforcement learning used in behavioral economics [22]. This learning algorithm is popular because it fits data from several experiments well and it generalizes several other learning algorithms. For a slightly simplified version of EWA [21] the strategy i for player Row evolves according to

where Π_i^R is the i -th row of the payoff matrix of player Row and $\Pi_i^R \mathbf{y}$ is the expected payoff for player Row when she plays move i and Column plays mixed strategy $\mathbf{y} = \{y_1, \dots, y_N\}$. The parameter β quantifies how strongly the payoffs are concentrated on the moves that have been most successful – when $\beta = 0$ all moves are equally likely, and when $\beta = \infty$ only the most successful move is used. The parameter α determines the relative weighting of recent vs. past performance, and $Z_x = \sum_{i=1}^N x_i(t+1)$ ensures proper normalization. A similar expression holds for y_j . At each time step, both players update the probabilities for all moves i and j .

$$x_i(t+1) = \frac{(x_i(t))^{1-\alpha} e^{\beta \Pi_i^R \mathbf{y}}}{Z_x}, \quad (1)$$

We also study evolutionary games in which x_i and y_j are the population shares of individuals with traits i and j respectively. This naturally leads to two-population replicator dynamics (RD),

$\dot{x}_i = x_i (\Pi_i^R \mathbf{y} - \mathbf{x} \Pi^R \mathbf{y})$,
 $\dot{y}_j = y_j (\Pi_j^C \mathbf{x} - \mathbf{y} \Pi^C \mathbf{x})$.

$$\dot{x}_i = x_i (\Pi_i^R \mathbf{y} - \mathbf{x} \Pi^R \mathbf{y}), \quad (2)$$

The shares of trait i in population Row and trait j in population Column evolve according to the fitness of that trait (as given by the expected payoff) compared to the average fitness in the respective population [11]. In fact it is easy to show that EWA and RD

are related [21, 23]. Note that while one-population RD often converges to mixed strategy Nash equilibria, two-population RD does not [3].

In this paper we assume deterministic dynamics, which results from batch learning (the players update their strategies only after observing an infinitely large sample of moves by their opponent) or from infinite populations. Finite sampling [24] and demographic noise [25] may have important effects, but in generic and possibly high-dimensional games stochasticity does not seem to play a major role [21].

We study the generic asymptotic properties of the dynamics above in an indirect manner. In particular, we study a third type of learning dynamics that we can understand analytically, and then show numerically that its generic properties are very similar to those of EWA or RD. This dynamics is based around the concept of a *best reply*, i.e. the move that gives the best payoff in response to a given move by an opponent. This is an old concept, originating in 1838 with Cournot [26]. Most criteria that have been used to assess convergence in game theory, such as dominance solvability and Evolutionary Stable States (ESS), are subsumed in the best reply structure. EWA and RD are more sophisticated and depend on *better replies*. The players do not only look at the most successful move, but also at all other moves that yield better payoffs than the move they are playing. However, we will show that best replies are what ultimately determines the probability of non-convergence.

We consider a particular version of *best reply dynamics* in which the two players alternate moves, each making her best response to her opponent's last move. To see the basic idea consider the game with $N = 4$ shown in Fig. 1A. Suppose we choose (1,1) as the initial condition. Assume Column moves first, choosing move $S^C = 2$, which is the best response to Row's move $S^R = 1$. Then Row's best response is $S^R = 2$, then Column moves $S^C = 1$, etc. This traps the players in the cycle (1,1) \rightarrow (1,2) \rightarrow (2,2) \rightarrow (2,1) \rightarrow (1,1), corresponding to the red arrows. We call this a 2-cycle, corresponding to the fact that each player moves twice. This cycle is an attractor, as can be seen by the fact that starting at (3,2) with a play by Row leads to the cycle. In general the first mover can be taken randomly; if the players are on a cycle, this makes no difference, but when off an attractor it can be important. In fact for this example there are two attractors: If Column had instead gone first, we would have instead arrived in one step at the fixed point at (3,3) (shown in blue). A fixed point of the best reply dynamics is called a *pure strategy Nash equilibrium*.

Our goal now is to compute the fraction \mathcal{F} of occurrences in which best reply dynamics do not converge

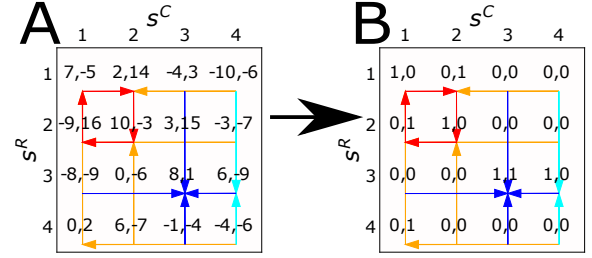


FIG. 1. *Illustration of best reply dynamics.* $S^R = \{1, 2, 3, 4\}$ and $S^C = \{1, 2, 3, 4\}$ are the possible moves of players Row and Column and each cell in the matrix represents their payoffs (Row is given first). The best response arrows point to the cell corresponding to the best reply. The vertical arrows correspond to player Row and the horizontal arrows to player Column. The arrows are colored red if they are part of a cycle, orange if they are not part of a cycle but lead to one, blue if they lead directly to a fixed point, and cyan if they lead to a fixed point in more than one step. The payoff matrix in B is a Boolean version that is constructed to have the same best reply structure as the payoff matrix in panel A.

to pure strategy Nash equilibria. We characterize the set of attractors of a given $N \times N$ payoff matrix Π by a vector $\mathbf{v}(\Pi) = (n_N, \dots, n_2, n_1)$, where n_1 is the number of fixed points, n_2 the number of 2-cycles, etc. For instance $\mathbf{v} = (0, 0, 1, 1)$ for the example in Fig. 1. We define $C = \sum_{k=2}^N n_k k$ as the number of moves that are part of cycles. The fraction of non-convergence of best reply dynamics is approximated by the size of the cycles vs. the fixed points, that is $\mathcal{F}(\mathbf{v}) = C/(C + n_1)$. In Fig. 1, $\mathcal{F}(0, 0, 1, 1) = 2/3$.

For any given matrix Π let the set of best replies by both players to all possible moves of their opponent be the *best reply configuration*. The total number of possible configurations is N^{2N} . Under the assumption that all Π are equally likely, we can compute the frequency $\rho(\mathbf{v})$ for any set of attractors \mathbf{v} according to the micro-canonical ensemble by counting the number of configurations leading to \mathbf{v} .

Here we just sketch the derivation, referring the reader to the Supplemental Information (SI) for a detailed explanation. Because of independence, the frequency $\rho(\mathbf{v})$ can be written as a product of terms corresponding to the number of ways to obtain each type of attractor, multiplied by a term h for best replies that are not on attractors. We denote by n the number of moves per player which are not already part of cycles or fixed points. The function $f(n, k)$ counts the ways to have a k -cycle, $g(n, k)$ counts the ways to have k distinct fixed points:

$$f(n, k) = \binom{n}{k}^2 k!(k-1)!, \quad (3)$$

$$g(n, k) = \binom{n}{k}^2 k!, \quad (4)$$

where the binomial coefficients mean that each player can choose any k moves out of n to form cycles or fixed points, and the factorials quantify all combinations of best replies that yield cycles or fixed points with the selected k moves. For instance, in Fig. 1, for each player we can choose any 2 moves out of 4 to form a 2-cycle, and for each of these choices there are two possible cycles (one going clockwise and the other counterclockwise). Similarly, for each player we can select any move out of the remaining two to form a fixed point. For both players we can still freely choose one best reply, provided this does not form another pure strategy Nash equilibrium. In Fig. 1, such best replies are (3, 4) for Row and (4, 1) for Column. In general, $h_N(n)$ counts the number of ways to combine the remaining n free best replies in a $N \times N$ payoff matrix, so that they do not form other cycles or fixed points:

$$h_N(n) = N^{2n} - \sum_{k=2}^n f_N(n, k, 0) - \sum_{k=1}^n g(n, k) h_N(n-k), \quad (5)$$

where the second term $\sum_{k=2}^n f_N(n, k, d) = \sum_{k=2}^n f(n, k) \left[N^{2(n-k)} - \sum_{j=2}^{n-k} \frac{f_N(n-k, j, d+1)}{d+2} \right]$ is used to count all possible ways for the n free best replies to form a cycle at recursion depth d . The division by $d+2$, like the division by j in Eq. (6), is needed to prevent double, triple, quadruple, etc. counting of cycles.

For any given set of attractors $\mathbf{v} = (n_N, \dots, n_2, n_1)$ the general expression for its frequency ρ is

$$\rho(\mathbf{v}) = \left(\prod_{k=2}^N \prod_{j=1}^{n_k} \frac{f\left(N - \sum_{l=k+1}^N n_l l - (j-1)k, k\right)}{j} \right) g\left(N - \sum_{l=2}^N n_l l, n_1\right) h_N\left(N - \sum_{l=2}^N n_l l - n_1\right) / (N^{2N}). \quad (6)$$

For the payoff matrix in Fig. 1, $\rho(0, 0, 1, 1) = f(4, 2)g(2, 1)h_4(1)/4^8 = 0.07$.

This can then be used to compute the average fraction of non-convergence of best reply dynamics \mathcal{F} for any given N ,

$$\mathcal{F}_N = \sum_{\mathbf{v}} \rho(\mathbf{v}) \mathcal{F}(\mathbf{v}), \quad (7)$$

summing over all possible \mathbf{v} s.t. $\sum_{k=1}^N n_k k \leq N$.

We now numerically test to see whether there is a correspondence between the fraction of situations where learning does not converge to a fixed point for EWA or RD and the frequency of non-convergence under best reply dynamics $\mathcal{F}(\mathbf{v})$, for specific best reply vectors or sets of attractors \mathbf{v} . We begin by

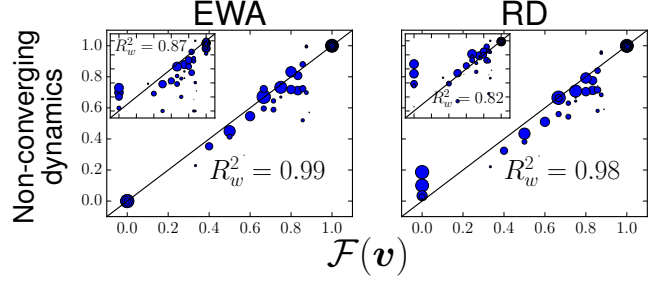


FIG. 2. Test for how well the best reply dynamics predicts non-convergence under EWA (left) and RD (right). The vertical axis gives the fraction of non-convergence in the simulations, and the position on the horizontal axis is the frequency of non-convergence under best reply dynamics $\mathcal{F}(\mathbf{v})$. Each dot corresponds to a specific best reply vector \mathbf{v} whose size is $\rho(\mathbf{v})$ from Eq. (6). In the main panels simulations are based on Boolean approximations of normally generated payoff matrices, whose simulations are shown in the insets. The identity line is plotted for reference.

performing Monte Carlo simulations with $N = 20$, generating 1000 random payoff matrices and testing 1000 randomly chosen initial conditions for each one. We measure the fraction of non-converging simulation runs and produce Fig. 2 using all observed \mathbf{v} . Payoff matrices with the same \mathbf{v} under best reply dynamics are grouped together. For EWA, for example, the large circle at (0,0) corresponds to $\mathbf{v} = (0, \dots, 0, x)$, the circle at (1,1) to $\mathbf{v} = (x, \dots, x, 0)$, and the large circle near $\mathcal{F}(\mathbf{v}) = 0.7$ corresponds to $\mathbf{v} = (0, \dots, 0, 1, 1)$. The details of the simulation protocol, the convergence criteria and the parameter values are listed in the SI. The main panels in Fig. 2 refer to simulation runs with Boolean approximations of the payoff matrices. In this case there is an extremely strong correlation between the simulations and the predicted values. The lack of perfect correlation is both due to non-linearities in the learning algorithms and to the positions of the best replies that are not on attractors, but may affect the respective basins of attraction.

The insets refer to simulation runs in the original (i.e. non-Boolean) payoff matrices. Other than the noise sources listed above, these are subject to an additional disturbance factor due to *quasi*-best replies, which occur when two payoffs are too close for the Boolean approximation to be valid. For instance, in Fig. 1A, $\Pi_{3,2}^C = 15$ and $\Pi_{1,2}^C = 16$: the two payoffs are very close and because of history dependence and limited rationality, player Column might switch to 3 rather than to 1 after the players chose (2, 2). As a consequence, the dynamics may converge to the pure strategy Nash equilibrium rather than remaining trapped in the best reply cycle. Despite these effects, the correlation is still very strong, with weighted cor-

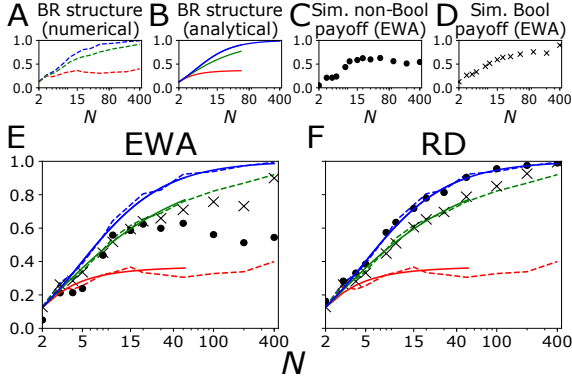


FIG. 3. A: The dashed lines depict the fraction of randomly generated payoff matrices with no pure strategy Nash equilibria (bottom red line, $\mathcal{F}(\mathbf{v}) = 1$), the fraction with at least one cycle (top blue line, $\mathcal{F}(\mathbf{v}) > 0$) and the frequency of non-convergence under best reply dynamics, averaged over all randomly generated payoff matrices (middle green line, \mathcal{F}_N). B: The solid lines have the same meaning, but are obtained from the analytical frequencies. C: Dots are the fraction of non-converging simulation runs. D: Crosses are the fraction of non-converging simulation runs in Boolean payoff matrices. E-F: Data in A-D superimposed, for EWA and RD respectively.

relation coefficients $R_w^2 = 0.87$ and $R_w^2 = 0.82$ for EWA and RD respectively. The weights are given by the frequency of the best reply vectors \mathbf{v} , because noise in the most common \mathbf{v} is averaged out.

In Fig. 3 we report the results from running Monte Carlo experiments for increasing values of N . Panel A shows the numerical fraction based on a best reply analysis of the random payoff matrices that were numerically generated and B contains the predicted fraction based on the analytical calculations of Eqs. (6) and (7). For instance, for $N = 50$, 36% of the payoff matrices have no fixed points, 10% have no cycles, and 54% have a mixture, with an average $\mathcal{F}_N = 0.76$. The larger the payoff matrix, the higher the probability of cycles. Fixed points do not disappear entirely but, as the middle green lines show, most moves are part of cycles. This is an important result: many papers in game theory focus on so-called dominance-solvable, coordination, potential and supermodular games [14–17], where learning generally converges. But all such games are best reply *acyclic*, so when N is large they only represent a small fraction of all possible payoff matrices. Likewise, ESS in two-population evolutionary games only correspond to pure strategy Nash equilibria [3], which become relatively rare as the payoff matrix grows larger.

Second, we consider the simulation runs. In Boolean payoff matrices, the fraction of non-converging simulations closely tracks the fraction \mathcal{F} , as observed in Fig. 2. In non-Boolean payoff matrices this is not always the case: we believe that

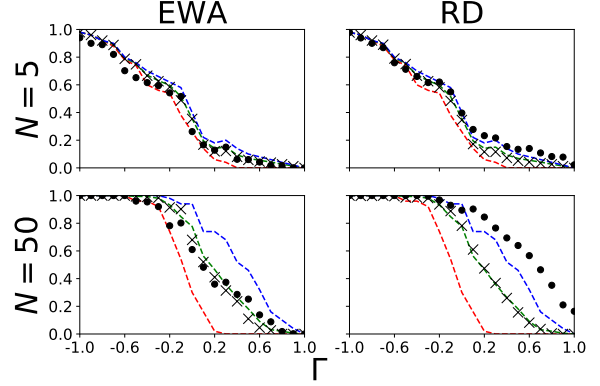


FIG. 4. The lines and symbols have the same meaning as in Fig. 3, but we constrain the average correlation Γ between the payoffs of the two players. A negative correlation increases the share of best reply cycles and hence the likelihood of non-convergence; a positive correlation has the opposite effect.

this effect is caused by quasi-best replies, as in Fig. 2 (insets). However, the fraction of non-converging simulation runs stays within the stability boundaries determined by the lower red and upper blue lines, indicating that theory has predictive value. Finally, we would like to stress that the reason why the RD simulations track the upper blue line and generally seem more unstable, as can also be noted in Fig. 2, is merely due to numerical limitations. We explain these in the SI.

Further evidence for the relevance of the best reply structure is provided in Fig. 4. Instead of picking all payoffs independently at random, we constrain the pairs of payoffs (i.e. the payoffs that Row and Column get when they play moves i and j) so that their average correlation [21] is Γ . Now best reply configurations are not equiprobable any more, so we cannot use the formalism in Eq. (6) and therefore we only proceed numerically, generating random payoff matrices and computing their best reply structure. A negative correlation, $\Gamma < 0$, implies that the game is competitive (zero-sum in the extreme case where $\Gamma = -1$), while $\Gamma > 0$ encourages cooperation, in the sense that combinations of moves tend to be either mutually beneficial or undesirable for both players. Intuitively this increases the chances for pure strategy Nash equilibria. On the contrary, competitive games are characterized by best reply cycles, and equilibria almost disappear if Γ is sufficiently negative. This causes non-converging dynamics in competitive games. Interestingly, if N is large, the behavior changes abruptly when Γ is varied, hinting at the existence of a continuous phase transition. Note that for large N and $\Gamma < 0$ the learning dynamics may settle into high-dimensional chaotic attractors [21].

In summary, we have proposed a new framework to understand the stability of learning algorithms and evolutionary dynamics that depend on generic payoff matrices. While it is well known that stylized processes such as best reply dynamics are influenced by the best reply structure of the payoff matrix, we have shown that the stability of highly non-trivial dynamics can be predicted to a good extent by the best reply structure. We have also shown that generic high dimensional payoff matrices rarely yield convergence because of the intertwined structure of the interactions they represent. The best reply dynamics gives useful insight into why non-convergence occurs that is not obvious from previous calculations in the large N limit based on path integrals.

We believe that the method introduced in this paper is of wider applicability. For instance, Lotka-Volterra equations are equivalent to the replicator dynamics [11], so we can use the best reply structure to assess stability of ecosystems [27]. It may be possible to relate these findings to the network properties of the food webs [28]. More generally, if equilibrium is not a good description of how a game is played, this questions the use of rational strategies to model agents in social and economic systems characterized by strategic interactions. An alternative could be simple behavioral rules and heuristics.

* marco.pangallo@maths.ox.ac.uk

- [1] Roger B Myerson. *Game theory*. Harvard university press, 2013.
- [2] John Maynard Smith. *Evolution and the Theory of Games*. Cambridge university press, 1982.
- [3] Herbert Gintis. *Game theory evolving: A problem-centered introduction to modeling strategic behavior*. Princeton university press, 2000.
- [4] Robert Axelrod and William D Hamilton. The evolution of cooperation. *Science*, 211(4489):1390–1396, 1981.
- [5] Martin A. Nowak and David C. Krakauer. The evolution of language. *Proceedings of the National Academy of Sciences*, 96(14):8028–8033, 1999.
- [6] Alessandro Di Mare and Vito Latora. Opinion formation models based on game theory. *International Journal of Modern Physics C*, 18(09):1377–1395, 2007.
- [7] Julia Robinson. An iterative method of solving a game. *Annals of mathematics*, pages 296–301, 1951.
- [8] G. W. Brown. Iterative solution of games by fictitious play. In T.C. Koopmans, editor, *Activity analysis of production and allocation*, pages 374–376. Wiley, New York, 1951.
- [9] Brian Skyrms. Chaos in game dynamics. *Journal of Logic, Language and Information*, 1(2):111–130, 1992.
- [10] Drew Fudenberg and David K Levine. *The theory of learning in games*, volume 2. MIT press, 1998.
- [11] Josef Hofbauer and Karl Sigmund. *Evolutionary games and population dynamics*. Cambridge university press, 1998.
- [12] Yuzuru Sato, Eizo Akiyama, and J Doyne Farmer. Chaos in learning a simple two-person game. *Proceedings of the National Academy of Sciences*, 99(7):4748–4751, 2002.
- [13] Daniele Vilone, Alberto Robledo, and Angel Sánchez. Chaos and unpredictability in evolutionary dynamics in discrete time. *Physical review letters*, 107(3):038101, 2011.
- [14] John H Nachbar. “evolutionary” selection dynamics in games: Convergence and limit properties. *International journal of game theory*, 19(1):59–89, 1990.
- [15] Dean P Foster and H Peyton Young. On the non-convergence of fictitious play in coordination games. *Games and Economic Behavior*, 25(1):79–96, 1998.
- [16] Dov Monderer and Lloyd S Shapley. Fictitious play property for games with identical interests. *Journal of economic theory*, 68(1):258–265, 1996.
- [17] Paul Milgrom and John Roberts. Adaptive and sophisticated learning in normal form games. *Games and economic Behavior*, 3(1):82–100, 1991.
- [18] Manfred Opper and Sigurd Diederich. Phase transition and $1/f$ noise in a game dynamical model. *Physical review letters*, 69(10):1616–1619, 1992.
- [19] J Berg and A Engel. Matrix games, mixed strategies, and statistical mechanics. *Physical Review Letters*, 81(22):4999–5002, 1998.
- [20] Johannes Berg. Statistical mechanics of random two-player games. *Physical Review E*, 61(3):2327–2339, 2000.
- [21] Tobias Galla and J Doyne Farmer. Complex dynamics in learning complicated games. *Proceedings of the National Academy of Sciences*, 110(4):1232–1236, 2013.
- [22] Colin Camerer and Teck Ho. Experience-weighted attraction learning in normal form games. *Econometrica*, 67(4):827–874, 1999.
- [23] Yuzuru Sato, Eizo Akiyama, and James P Crutchfield. Stability and diversity in collective adaptation. *Physica D: Nonlinear Phenomena*, 210(1):21–57, 2005.
- [24] Tobias Galla. Intrinsic noise in game dynamical learning. *Physical review letters*, 103(19):198702, 2009.
- [25] Alan J McKane and Timothy J Newman. Predator-prey cycles from resonant amplification of demographic stochasticity. *Physical review letters*, 94(21):218102, 2005.
- [26] Antoine-Augustin Cournot. *Recherches sur les principes mathématiques de la théorie des richesses*. L. Hachette, 1838.
- [27] Robert M May. Will a large complex system be stable? *Nature*, 238:413–414, 1972.
- [28] Samuel Johnson, Virginia Domnguez-Garca, Luca Donetti, and Miguel A. Muoz. Trophic coherence determines food-web stability. *Proceedings of the National Academy of Sciences*, 111(50):17923–17928, 2014.

Supplemental information for:
Best reply structure and equilibrium convergence in generic games

DETAILS OF THE ANALYTICAL CALCULATIONS

We explain the derivation of Eq. (6) in the main paper in more detail. We then provide an example of Eq. (6) in its full generality, that is we calculate the frequency of a specific best reply vector (or set of attractors) with $N = 11$. We finally present two complimentary results that we did not mention in the main paper: we list the most common best reply vectors for $N = 11$ and $N = 20$ and we obtain an asymptotic estimate for the frequency of k -cycles in infinite dimensional payoff matrices.

Frequency of best reply vectors

We first discuss the count of the ways to form k -cycles and fixed points of best reply dynamics, and to place the “free” best replies (i.e. those that are not part of either cycles or fixed points). Finally we show how we combine these numbers together to obtain the count of best reply configurations that correspond to a specific set of attractors.

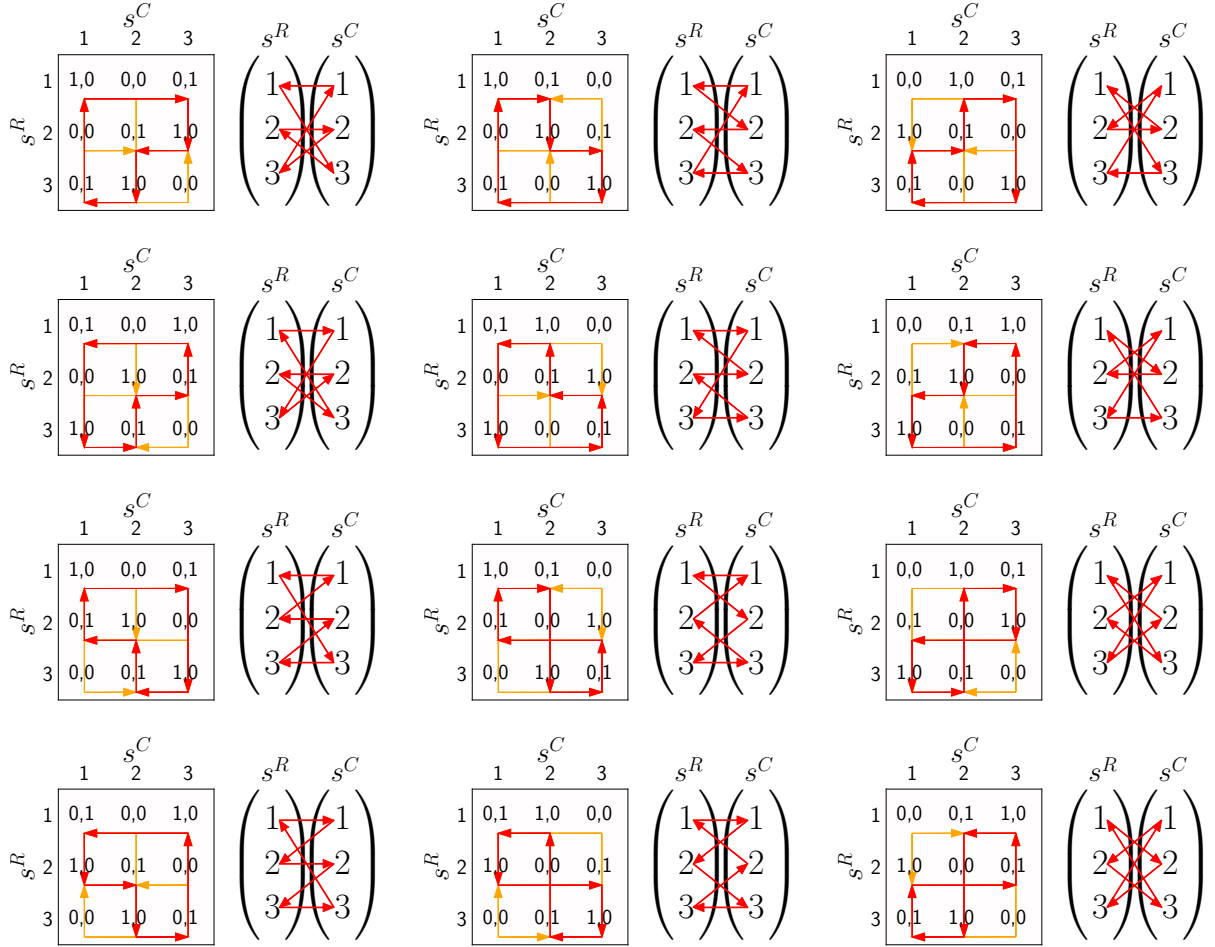


FIG. S1. All possible $3! \cdot 2! = 12$ ways to combine 3 moves per player to form 3-cycles. The color code has been kept consistent with the main text. The (1,2,3) vertical arrays contain the labels of the moves and the arrows represent the best replies. A cycle is a closed loop of best replies. These 12 combinations are also all best reply configurations featuring a 3-cycle in payoff matrices with $N = 3$. Using Eq. (S1), $f(3,3) = 12$.

We start the count of k -cycles by example. In Fig. S1 we exhaustively report all possible ways to form 3-cycles in a payoff matrix with $N = 3$. The vertical $(1, 2, 3)$ arrays and the arrows that connect the labels of the moves illustrate the main intuition: we find all possible best reply sequences that form a closed loop. We arbitrarily start at $s^R = 1$ (because this is a cycle, the starting point does not matter), we look at the best reply by player Column, $s^C \in \{1, 2, 3\}$, and we connect $s^R = 1$ with s^C . In the top left panel, we connect $s^R = 1$ to $s^C = 3$. The first choice can be done in $k = 3$ ways. Once we have determined the first best reply by Column, we continue constructing the cycle by choosing a second best reply by Row. The second choice can only be done in $k - 1 = 2$ ways. In the top left panel, we connect $s^C = 3$ to $s^R = 2$. We then select a second best reply by Column. Again, we have $k - 1 = 2$ possibilities. In the top left panel, we connect $s^R = 2$ to $s^C = 2$. The third and last best replies for Row and Column are constrained, there is only one ($k - 2 = 1$) way to choose the remaining BR. In the top left panel, we connect $s^C = 2$ to $s^R = 3$ and $s^R = 3$ to $s^C = 1$. We have $3 \cdot 2 \cdot 2 \cdot 1 \cdot 1 = 12$ ways to form 3-cycles with $n = 3$ available moves. Recall that n denotes the number of moves per player which are not already part of cycles or fixed points. In general n might be smaller than N , but in Fig. S1 all moves are part of the cycle, so $N = n = k = 3$.

It is possible to generalize this argument and to conclude that there are $k!(k - 1)!$ ways to form k -cycles, once we determine which moves of players Row and Column are involved. Any k moves out of n can be chosen (by both players), so there are $\binom{n}{k}^2$ possibilities. We define

$$f(n, k) = \binom{n}{k}^2 k!(k - 1)!, \quad (\text{S1})$$

with $2 \leq k \leq n$, as the count of the ways to have a k -cycle with n available moves per player. In the above example, $f(3, 3) = 12$.

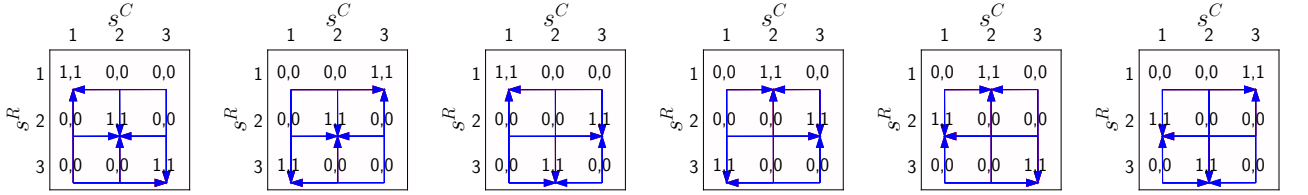


FIG. S2. All possible $3! = 6$ ways to combine 3 moves per player to form 3 fixed points. The color code has been kept consistent with the main text. Note that these are also all best reply configurations featuring 3 fixed points in payoff matrices with $N = 3$. Using Eq. (S2), $g(3, 3) = 6$.

We now look at the ways to form fixed points, and we begin again by example. In Fig. S2 we report all possible ways to form 3 fixed points in a payoff matrix with $N = 3$. Once we determine which moves are part of the fixed points (all, in this case), we form all possible combinations of fixed points by picking pairs of moves from the lists of available moves by both players. For convenience, we start again from $s^R = 1$. We form a fixed point by choosing any move $s^C \in \{1, 2, 3\}$, so that (s^R, s^C) is a fixed point. In the left panel, we choose $(1, 1)$ as the first fixed point. We then consider $s^R = 2$. There are only two moves available from player Column to form a second fixed point. In the left panel, $(2, 2)$ is the second fixed point. Finally, for $s^R = 3$ only one move by Column is available. By process of elimination, in the left panel $(3, 3)$ is the third and last fixed point.

In general, there are $k!$ ways to form k fixed points once k moves are chosen. The count of the ways to select the k moves out of n is identical with respect to the cycles, so we define

$$g(n, k) = \binom{n}{k}^2 k!, \quad (\text{S2})$$

with $1 \leq k \leq n$, as the count of the ways to have k fixed points with n available moves per player. In the above example, $g(3, 3) = 6$.

We finally calculate the ways to place the free best replies, which are not part of either cycles or fixed points. We begin again by example. In Fig. S3 we show payoff matrices with one free best reply per player. In the top left panel, the best reply of Row to Column playing $s^C = 3$ is $s^R = 2$; the best reply of Column to Row playing $s^R = 3$ is $s^C = 3$. The free best replies can be chosen freely, except for both of them to be move 3, in

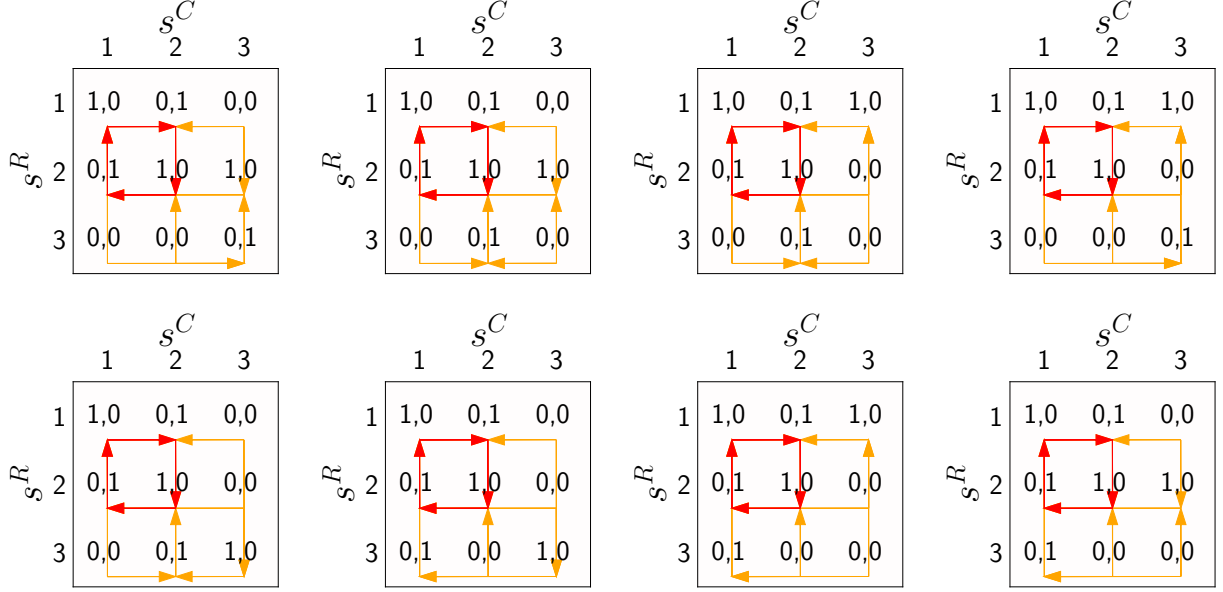


FIG. S3. All possible $3^2 - 1 = 8$ ways to choose the two remaining best replies, so that they do not form a fixed point at $(3, 3)$. The color code has been kept consistent with the main text. Using Eq. (S3), $h_3(1) = 8$.

which case they would form another fixed point. In this example there are $3^2 - 1 = 8$ ways to choose free best replies so that they do not form other cycles or fixed points.

In general,

$$h_N(n) = N^{2n} - \sum_{k=2}^n f_N(n, k, 0) - \sum_{k=1}^n g(n, k) h_N(n - k) \quad (\text{S3})$$

counts all possible ways to combine n free best replies in a $N \times N$ payoff matrix, so that they do not form other cycles or fixed points. We provide a more complete example for Eq. (S3) at the end of this section. Note that N is a parameter and therefore is indicated as a subscript, while n is a recursion variable: even when the number of available moves n is smaller than N , the free best replies can be chosen out of all the N moves (see Fig. S3), in N^{2n} ways. In Eq. (S3),

$$\sum_{k=2}^n f_N(n, k, d) = \sum_{k=2}^n f(n, k) \left[N^{2(n-k)} - \sum_{j=2}^{n-k} \frac{f_N(n-k, j, d+1)}{d+2} \right] \quad (\text{S4})$$

is used to count all possible ways for the n free best replies to form a cycle, at recursion depth d . Every term k in the summation is the number of k -cycles, $f(n, k)$, multiplied by all possible $N^{2(n-k)}$ ways to place the $n - k$ remaining best replies by both players, minus a term $\sum_{j=2}^{n-k} f_N(n-k, j, d+1)/(d+2)$ that excludes from the count the cycles that can be formed with the $n - k$ remaining best replies. The division by $d+2$ is necessary to avoid double, triple, quadruple, ... $(d+2)$ -ple counting of cycles. Consider for instance the calculation of the number of 2-cycles in a 4×4 payoff matrix: $N = n = 4, k = 2, d = 0$. By using the formalism above, $f_4(4, 2, 0) = f(4, 2) [4^{2-2} - f_4(2, 2, 1)/2]$, where $f_4(2, 2, 1) = f(2, 2) [4^0] = 2$. There is a number $f(4, 2)$ of 2-cycles, and for each of these there are 4^4 ways to place the two remaining best replies of the players. But if those are combined so that they form another 2-cycle, we would count 2-cycles twice, so we need to remove one best reply configuration from the count. Moreover, in Eq. (S3),

$$\sum_{k=1}^n g(n, k) h_N(n - k) \quad (\text{S5})$$

counts all possible ways for the n free best replies to form at least one fixed point.

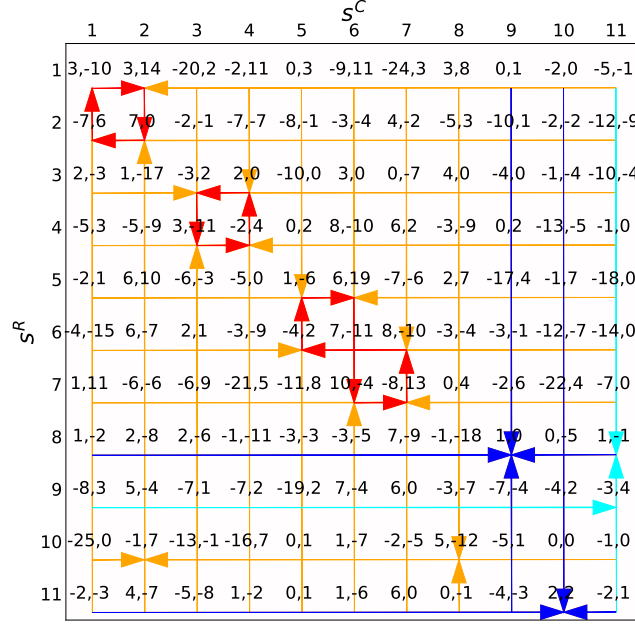


FIG. S4. Payoff matrix with $N = 11$. The color code has been kept consistent with the main text. The set of attractors of best reply dynamics in the payoff matrix is $\mathbf{v} = (0, 0, 0, 0, 0, 0, 0, 0, 1, 2, 2)$, with $n_3 = 1$, $n_2 = 2$, $n_1 = 2$ and $n_k = 0$, if $k > 3$. It is $\sum_{k=1}^{11} n_k k = 9 < 11$.

We now combine all the ways to have cycles, fixed points and free best replies to calculate the number of best reply configurations that correspond to a generic best reply vector $\mathbf{v} = (n_N, n_{N-1}, \dots, n_k, \dots, n_2, n_1)$. We denote by n_1 the number of fixed points and by n_k , with $2 \leq k \leq N$, the number of k -cycles. Of course \mathbf{v} has to obey the obvious constraint that fixed points and k -cycles do not take up more than N moves: $\sum_{k=1}^N n_k k \leq N$. The frequency of the best reply vector \mathbf{v} is

$$\rho(\mathbf{v}) = \left(\prod_{k=2}^N \prod_{j=1}^{n_k} \frac{f\left(N - \sum_{l=k+1}^N n_l l - (j-1)k, k\right)}{j} \right) g\left(N - \sum_{l=2}^N n_l l, n_1\right) h_N\left(N - \sum_{l=2}^N n_l l - n_1\right) / (N^{2N}). \quad (\text{S6})$$

Eq. (S6) is Eq. (6) in the main paper. The first term with f counts all the ways to have k -cycles, by multiplying the counts for all values of k (first product) and for all k -cycles for a specific value of k (second product). Note that we progressively reduce the number of moves available to form k -cycles, as more and more moves become part of k -cycles (see below for an example that clarifies this point). If there are multiple k -cycles, $n_k > 1$, we divide the count by $j = 1, \dots, n_k$ so to avoid double, triple, etc. counting. The second term g counts all the ways to have n_1 distinct fixed points within the remaining $N - \sum_{l=2}^N n_l l$ moves. The third term h counts all the ways to choose the remaining $N - \sum_{l=2}^N n_l l - n_1$ free best replies. The product of the three terms gives the number of best reply configurations that correspond to the best reply vector \mathbf{v} . We divide this number by the possible configurations N^{2N} and we obtain the frequency.

As an example, we calculate the number of best reply configurations with the same set of attractors as in Fig. S4. First, we can choose any 3 moves out of 11 for both players to be part of a 3-cycle, meaning that there are $\binom{11}{3}^2$ possibilities. Second, we can obtain 12 cycles for each choice of 3 moves per player by choosing $3!2! = 12$ sequences of moves. The same reasoning applies to the two 2-cycles, except that there are only 8 and 6 moves per player still available and that the count of the ways to have 2-cycles needs to be divided by 2 in order to avoid double counting. The number of best reply configurations with 2 fixed points in the remaining 4 moves can be calculated similarly: each player can choose any 2 moves out of 4, and then there are 2 ways to select the pairings. We are left with 2 moves that are not part of cycles or fixed points. There are 11^4 ways to choose the free best replies, but we have to exclude the cases in which they would form another 2-cycle or one or more fixed points. There are 2 ways they could form a 2-cycle, 2 ways they could form 2 fixed points and 4 ways they could form 1 fixed point. But for each of the latter there are 11^2 ways to choose the free best replies, minus the way in which this choice would form another fixed point (which had already been counted

in the configurations with 2 fixed points). In summary, the number of best reply configurations is given by

$$\rho(0, 0, 0, 0, 0, 0, 0, 0, 1, 2, 2) = f(11, 3)f(8, 2)\frac{f(6, 2)}{2}g(4, 2)h_{11}(2)/(11^{22}), \quad (S7)$$

with $f(11, 3) = \binom{11}{3}^2 3 \cdot 2 \cdot 2$, $f(8, 2) = \binom{8}{2}^2 2$, $f(6, 2) = \binom{6}{2}^2 2$, $g(4, 2) = \binom{4}{2}^2 2$ and $h_{11}(2) = 11^4 - 2 - 2 - 4 \cdot (11^2 - 1)$.

The explicit computation of the frequency gives $\rho(0, 0, 0, 0, 0, 0, 0, 0, 1, 2, 2) = 1.44 \cdot 10^{-6}$, so the best reply vector in Fig. S4 is very infrequent. For $N = 11$, the most common best reply vectors are:

$$\begin{aligned} \rho(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1) &= 0.17, \\ \rho(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2) &= 0.14, \\ \rho(0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0) &= 0.14, \\ \rho(0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1) &= 0.13, \\ \rho(0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0) &= 0.09. \end{aligned} \quad (S8)$$

For $N = 20$, the most common best reply vectors are:

$$\begin{aligned} \rho(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1) &= 0.10, \\ \rho(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1) &= 0.10, \\ \rho(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2) &= 0.09, \\ \rho(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0) &= 0.09, \\ \rho(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0) &= 0.07. \end{aligned} \quad (S9)$$

We observe that k -cycles with high values of k are never really frequent; the frequency of any specific best reply vector decreases with N (because there are many more best reply vectors with positive frequency); the best reply vectors with cycles become more frequent as N increases, consistently with Fig. 3 of the main paper. Note that an accurate numerical estimate of the most common best reply vectors might be challenging due to the extremely high number of best reply configurations: the analytical result makes it possible to obtain exact estimates.

We also mention a technical detail concerning Fig. 3 of the main paper: the analytical lines for the fraction of payoff matrices with no fixed points (bottom red line, $\mathcal{F}(\mathbf{v}) = 1$), and with the frequency of non-convergence under best reply dynamics averaged over all best reply configurations (middle green line, \mathcal{F}) stop at $N = 50$, while the analytical line for the fraction of payoff matrices with at least one cycle (top blue line, $\mathcal{F}(\mathbf{v}) > 0$) continues up to $N = 400$. This is due to the fact that to compute the bottom and middle lines we need to explicitly calculate the frequency of all best reply vectors, whereas to compute the top line it is enough to use Eq. (S4), which is much less expensive computationally.

Absolute and relative frequencies of cycles

We define

$$f'_N(n, k, d) = f(n, k) \left[N^{2(n-k)} - \frac{f'_N(n-k, k, d+1)}{d+2} \right]. \quad (S10)$$

This expression is analogous to Eq. (S4), but it only considers double counting of k -cycles, and not of all j -cycles, $j = 2, \dots, N$. Therefore, $f'_N(N, k, 0)$ is the number of payoff matrices having at least one k -cycle. Note that $\sum_{k=2}^N f'_N(N, k, 0)$ sums to more than N^{2N} , because several best reply configurations have several k -cycles, with more than one value of k .

The usefulness of Eq. (S10) is that it can be used, in the limit $N \rightarrow \infty$, to calculate analytically the absolute and relative frequencies of k -cycles. We make the ansatz that the frequency of k -cycles reaches a fixed point when $N \rightarrow \infty$: $f'_N(N, k, 0)/(N^{2N}) \approx f'_N(N-k, k, 0)/(N-k)^{2(N-k)}$. This ansatz is reasonable at least for small k .

We then divide Eq. (S10) by N^{2N} :

$$\frac{f'_N(N, k, 0)}{N^{2N}} = \frac{N^2(N-1)^2 \dots (N-k+1)^2}{(k!)^2} k!(k-1)! \frac{\left[N^{2(N-k)} - \frac{f'_N(N-k, k, 1)}{2} \right]}{N^{2N}}. \quad (S11)$$

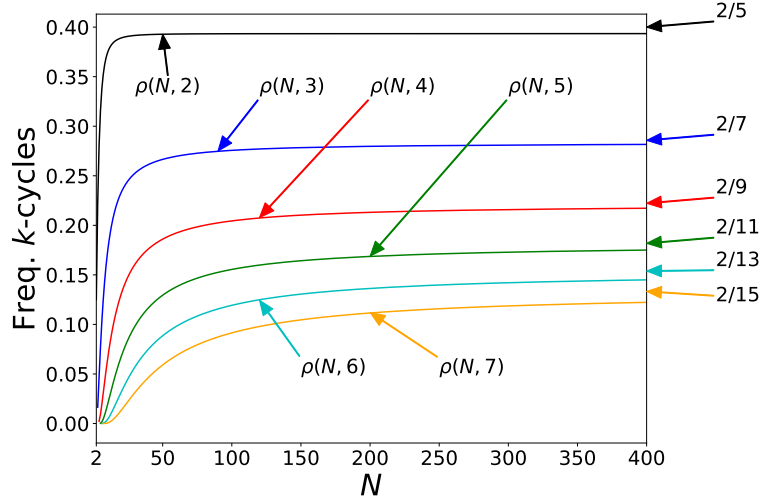


FIG. S5. Frequency of k -cycles, $\rho(N, k) = f'_N(N, k, 0)/N^{2N}$, as a function of the number of moves N . The numbers annotated on the right are the asymptotic frequencies of k -cycles, as calculated using Eq. (S13). The approximations tend to slightly overestimate the frequency, at least up to $N = 400$, even more so for larger values of k .

By applying the above ansatz and after some algebra we obtain

$$\lim_{N \rightarrow \infty} \frac{f'_N(N, k, 0)}{N^{2N}} := \rho(k) = \frac{1}{(k!)^2} k!(k-1)!(1 - \rho(k)/2), \quad (\text{S12})$$

which can be solved self-consistently to yield

$$\rho(k) = \frac{2}{2k+1}. \quad (\text{S13})$$

So for $N \rightarrow \infty$, 2-cycles appear in $2/5$ of the payoff matrices, 3-cycles in $2/7$, 4-cycles in $2/9$, etc. From Eq. (S13) we can easily obtain the relative frequencies:

$$\frac{f(k)}{f(2)} = \frac{5}{2k+1}, \quad (\text{S14})$$

so 3-cycles appear $5/7$ as often as 2-cycles, 4-cycles $5/9$ as often, 5-cycles $5/11$ as often, etc.

In Fig. S5 we report the frequency of k -cycles, as calculated using Eq. (S11), as a function of the number of available moves N . There is a good correspondence between the asymptotic behavior in Eq. (S13) and the explicit computation up to $N = 400$, at least for the smallest values of k .

DETAILS OF THE SIMULATION PROTOCOL

Experience-Weighted Attraction learning algorithm

Simulations of Experience-Weighted Attraction (EWA) are relatively straightforward. It is enough to iterate the map defined in Eq. (1) of the main paper, for each move i and j and starting from initial conditions $\mathbf{x}(0)$ and $\mathbf{y}(0)$.

EWA has two main advantages from a computational point of view. First, if the parameter which determines the relative weighting of recent vs. past performance, or memory loss parameter, is positive, $\alpha > 0$, all stable attractors of the dynamical system lie *within* the probability simplex. This means that no moves are ever given null or unit probability and makes it possible to reliably simulate the EWA map for arbitrarily long time, since for a sufficiently large value of α the numeric limits of the computer (10^{-316} with the *Python* programming language that we use) are never reached. The intuition for this property is simple: the performance of very successful or very unsuccessful moves is forgotten exponentially over time, so even a very small value of α

prompts the players to choose unsuccessful moves with positive probability. This property can be proved rigorously by evaluating the Jacobian of the EWA map at the vertices of the probability simplex, which are always fixed points of the dynamics: the eigenvalues are infinite. In Ref. [S1] we showed this for 2×2 games, i.e. games with $N = 2$ moves per player and 2 players. The second advantage is that the EWA system is explicitly normalized every time step, making numerical errors unlikely.

There is also a computational disadvantage: because EWA uses exponential functions to map payoffs into probabilities, if the value of the parameter which quantifies how strongly the payoffs are concentrated on the moves that have been most successful, or payoff sensitivity parameter β , is too large, the components of the mixed strategy vector may vary by too many orders of magnitude, and therefore overshoot the numeric limits of the computer.

So care should be taken in choosing the values of α and β . This is the case also because of an additional feature of the EWA system: with large memory loss or small payoff sensitivity, the learning dynamics converges to the center of the strategy simplex. In the limit where $\beta = 0$ the players just choose uniformly at random between their possible moves, irrespective of the payoff matrix. In Ref. [S2] it was observed that for sufficiently large values of α/β a unique fixed point was always stable. Such a fixed point could be characterized statistically using path integral methods [S3], and typically all components of the mixed strategy vector were of the same order of magnitude. Note that the fixed point can be arbitrarily far from mixed strategy Nash equilibria, and so by changing their strategy the players can improve their payoff. We are not interested in this “trivial” attractor as we want to focus on the effect of the best reply structure of the payoff matrix on the learning dynamics. Therefore, we choose parameter values for α and β that prevent convergence to this fixed point.

A final important technical remark is that we rescale the payoff sensitivity β by \sqrt{N} as the payoff matrix gets larger. The reason is that the expected payoffs $\Pi_i^R \mathbf{y}$ and $\Pi_j^C \mathbf{x}$ scale as $1/\sqrt{N}$. Indeed, for the expected payoff of player Row, $\sum_j \Pi_{ij}^R$ scales as \sqrt{N} due to the Central Limit Theorem (recall that the payoffs are generated randomly, see below for the precise rule), while the components y_j scale as $1/N$ due to the normalization constraint. So $\Pi_i^R \mathbf{y} = \sum_j \Pi_{ij}^R y_j$ scales as $1/\sqrt{N}$. The same argument applies to the expected payoff of player Column. Therefore, increasing the size of the payoff matrix has the same effect as decreasing β , until the attractor at the center of the strategy simplex becomes stable again. To prevent this from happening, we rescale β by \sqrt{N} , so that $\beta \Pi_i^R \mathbf{y}$ and $\beta \Pi_j^C \mathbf{x}$ do not scale with N .

Replicator Dynamics

Simulations of Replicator Dynamics (RD) are technically more challenging. First of all, RD is a set of differential equations that need to be discretized in order to be simulated. We use the Euler discretization

$$\begin{aligned} x_i(t+1) &= x_i(t) + x_i(t)\delta t \left(\Pi_i^R \mathbf{y} - \mathbf{x} \Pi^R \mathbf{y} \right), \\ y_j(t+1) &= y_j(t) + y_j(t)\delta t \left(\Pi_j^C \mathbf{x} - \mathbf{y} \Pi^C \mathbf{x} \right), \end{aligned} \tag{S15}$$

where δt is the integration step and should be chosen small enough to prevent overshooting of the boundaries of the probability simplex.

There are two main technical problems. First, in two-population RD, only *strict* Nash Equilibria are Evolutionary Stable States (ESS), that is (locally) stable fixed points of RD [S4]. No mixed strategy Nash equilibria are strict, meaning that they are never an attractor of the dynamics. On the contrary, all pure strategy Nash equilibria in randomly generated games are strict, because it never occurs that two payoffs are identical. Therefore all stable fixed points sit at the boundaries of the probability simplex. Moreover, because the RD system has infinite memory, cycles are not periodic. On the contrary, their period increases exponentially over time and the dynamics unavoidably drifts towards the edges of the probability simplex. So the map (S15) can be reliably simulated only for a limited *confidence time interval*: we stop the simulation run as soon as one component of \mathbf{x} or \mathbf{y} reaches the numeric limits of the computer. This is necessary because, if the dynamics is following a cycle, a certain move may not be played for a long time interval, with its probability decreasing over time. At some point, it may become convenient for the player to choose that move again, so the probability would start increasing again. But if the probability had hit the numeric limits of the computer beforehand, it would be stuck at zero, falsely identifying the simulation run as having reached a fixed point.

The second problem is that rounding approximations imply that normalization may be lost. If that happens, we stop the simulation run and discard the results.

Initialization of the payoff matrices

In order to study generic payoff matrices, we sample the space of all possible payoff matrices by generating the payoff elements at random. Following Ref. [S2], at initialization we randomly generate N^2 pairs of payoffs (i.e., if Row plays i and Column plays j , a pair a, b implies that Row gets a , Column gets b), and we keep the payoff matrix fixed for the rest of the simulation (so the system described by the payoff matrix can be thought of as *quenched*). We consider an *ensemble* of payoff matrices constrained by the mean, variance and correlation of the pairs. The Maximum Entropy distribution that obeys these constraints is a bivariate Gaussian [S2], which we parametrize with zero mean, unit variance and correlation Γ . Therefore, $\Gamma < 0$ implies that the game is competitive (zero-sum in the extreme case where $\Gamma = -1$), while $\Gamma > 0$ encourages cooperation (see the main text). If $\Gamma = 0$ all best reply configurations are equiprobable because the payoffs are chosen *independently* at random, so we shall consider this as a benchmark case where we sample the space of all possible games with equal probability.

Convergence criteria

The convergence criteria are different for EWA and RD:

- EWA: In order to be sure to identify the long-term behavior, we make a conservative choice for the length of the simulations. We run the EWA dynamics for 50000 time steps and we record $\mathbf{x}(t)$ and $\mathbf{y}(t)$ during the last 10000 time steps. With the parameter values we choose for α and β , the transient is usually of the order of 100 time steps, but we occasionally observe transient chaos [S2, S5] that lasts longer. We then check that the average variance of the logarithms of the components of the mixed strategy vectors does not exceed a certain (very small) threshold. We look at the logarithms because the probabilities following the EWA dynamics vary on an exponential scale and can be of the order of 10^{-100} , so we want to be sure that we do not identify as convergent a simulation run where the amplitude of a cycle is small. In formula, if $1/N \sum_{i=1}^N 1/T \sum_{t=4/5T}^T (\log x_i(t))^2 > 10^{-4}$ or $1/N \sum_{j=1}^N 1/T \sum_{t=4/5T}^T (\log y_j(t))^2 > 10^{-4}$, with $T = 50000$, we identify the simulation run as non-convergent.
- RD: With the integration step we choose, the confidence time interval is typically of the order of 1000 time steps. Because two-population RD cannot reach attractors within the probability simplex, the concept of transient is not well defined, so we cannot use the same convergence criterion as for EWA. Moreover, because the period of the cycles increases exponentially over time, if we look at the variance of the probabilities during the final time steps of the confidence time interval, we may falsely conclude that the dynamics had reached a fixed point. Likewise, if a pure strategy Nash equilibrium is approached at the very end of the confidence time interval, we may falsely identify the dynamics as non-converging. We propose the following heuristic: we check whether in the last 1/5 of the time steps the probabilities of the most used moves for both players are monotonically increasing, while all other probabilities are monotonically decreasing. Indeed, this is what we observe when the RD approaches a pure strategy Nash equilibrium, whereas there is a turning point when the RD is asymptotically trapped into a cycle. While we cannot conclude that this heuristic works in general, a direct inspection of over 100 simulation runs for several values of N confirms that convergence to pure strategy Nash equilibria or to best reply cycles has always been correctly identified.

Parameter values in simulation runs

For all simulations we choose $\alpha = 0.18$ and $\beta = \sqrt{N}$, which ensure that the EWA dynamics stays within the probability simplex, that it does not overshoot the simplex boundaries and that it does not reach the trivial attractor in the center of the simplex. We simulate the RD by choosing an integration step of $\delta t = 0.1$, and using the precautions described above.

Fig. 2 of the main paper: We generate 1000 payoff matrices at random with $\Gamma = 0$ and $N = 20$, starting from 1000 random initial conditions for each payoff matrix.

Fig. 3 of the main paper: We generate 180 payoff matrices at random with $\Gamma = 0$, starting from 10 random initial conditions for each payoff matrix, for the following numbers of moves: $N = \{2, 3, 4, 5, 8, 10, 15, 20, 30, 50, 100, 200, 400\}$. We sensibly reduce the number of simulation runs per value of N because the random

generation of the payoff matrix, the identification of the best reply structure and the simulations of the dynamics are time consuming for $N \geq 50$.

Fig. 4 of the main paper: Same as in Fig. 3, but we consider correlations $\Gamma = \{-1.0, -0.9, -0.8, \dots, 0.0, 0.1, \dots, 0.9, 1.0\}$ and only 50 payoff matrices. We consider $N = 5$ and $N = 50$ moves.

Finally, we would like to add a word of caution on the seemingly stronger instability of RD as compared to EWA, as it is seen in Figs. 2-4. Because of infinite memory and depending on the initial condition, it might take long to “find” a pure strategy Nash equilibrium, meaning that the RD can hit the numeric limits of the computer first, when it is still in a “transient”. In other words, it may not be in the basin of attraction determined by a cycle, but it may also have not reached a pure strategy Nash equilibrium within the confidence time interval. This is especially the case for large payoff matrices, $N \geq 50$. Indeed, note that for $N = 400$ almost all RD simulation runs are identified as non-converging, departing from the trend valid up to $N = 200$. We believe that this effect is just due to the numerical constraints described above.

* marco.pangallo@maths.ox.ac.uk

- [S1] Marco Pangallo, James BT Sanders, Tobias Galla, and J Doyne Farmer. A taxonomy of learning dynamics in 2 x 2 games. Preprint available at <https://arxiv.org/abs/1701.09043>, 2017.
- [S2] Tobias Galla and J Doyne Farmer. Complex dynamics in learning complicated games. *Proceedings of the National Academy of Sciences*, 110(4):1232–1236, 2013.
- [S3] Manfred Oppen and Sigurd Diederich. Phase transition and 1/f noise in a game dynamical model. *Physical review letters*, 69(10):1616–1619, 1992.
- [S4] Herbert Gintis. *Game theory evolving: A problem-centered introduction to modeling strategic behavior*. Princeton university press, 2000.
- [S5] James BT Sanders, Tobias Galla, and J Doyne Farmer. The prevalence of complex dynamics in games with many players. Preprint available at <https://arxiv.org/abs/1612.08111>, 2016.